DOCUMENT RESUME

ED 243 949 TM 840 264

AUTHOR Cziko, Gary A.; Lin, Nien-Hsuan Jennifer
TITLE The Construction and Analysis of Short Scales of Language Proficiency: Classical Psychometric, Latent Trait, and Nonparametric Approaches.
PUB DATE Apr 84
NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adults; *English (Second Language); Foreign Students; Higher Education; Item Analysis; *Language Proficiency; Latent Trait Theory; Measurement Techniques; Nonparametric Statistics; *Rating Scales; *Test Construction; Test Theory
IDENTIFIERS Illinois English Placement Test; Test of English as a Foreign Language

ABSTRACT

This study used classical psychometric, latent trait, and nonparametric approaches to analyze 13- and 14-item scales of English language proficiency. Tests of English listening comprehension (dictation) and reading ("copytest") were constructed by modifying the standard dictation testing procedure to create items of text segments which varied widely in both length and difficulty. Both the dictation and copytest were found to be homogeneous, cumulative scales of language proficiency with high reliability and validity. Log ability scores provided by Rasch analyses were found to correlate better with other measures of language proficiency than did the dictation and copytest raw scores. These findings indicate that the two language testing techniques investigated provide a useful innovative approach to measuring general aspects of language proficiency. The theoretical and practical advantages of this approach over other language proficiency measurement techniques are discussed as well as implications for measuring language proficiency and other cognitive variables. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED243949

The Construction and Analysis of Short Scales of Language Proficiency: Classical Psychometric, Latent Trait, and Nonparametric Approaches

Gary A. Cziko and Nien-Hsuan Jennifer Lin
University of Illinois at Urbana-Champaign

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. Cziko

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 840 269

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 1984.

Abstract

This study used classical psychometric, latent trait, and nonparametric approaches to analyze 13- and 14-item scales of English language proficiency. Tests of English listening comprehension (dictation) and reading ("copytest") were constructed by modifying the standard dictation testing procedure to create items of text segments which varied widely in both length and difficulty. Both the dictation and copytest were found to be homogeneous, cumulative scales of language proficiency with high reliability and validity. Log ability scores provided by Rasch analyses were found to correlate better with other measures of language proficiency than did the dictation and copytest raw scores. These findings indicate that the two language testing techniques investigated provide a useful innovative approach to measuring general aspects of language proficiency. The theoretical and practical advantages of this approach over other language proficiency measurement techniques are discussed as well as implications for measuring language proficiency and other cognitive variables.

The Construction and Analysis of Short Scales of Language Proficiency: Classical Psychometric, Latent Trait, and Nonparametric Approaches

Both the theory and practice of assessing second language proficiency has undergone marked changes over the last 40 years. Spolsky (1978) has classified language testing theory and practices into three major trends or periods to characterize the major changes which have taken place in the field. The first "prescientific" period relied primarily on teachers' subjective assessments of their students' ability to speak and/or write the language. During this period, generally before the 1960s in the U. S., there was little concern with the statistical reliability or validity of language assessment but rather an assumption that anyone proficient enough to teach a language would be also qualified to assess students' proficiency in it. The publication of Lado's Language Testing in 1961 marked the beginning of a second era in language testing, the "psychometric-structuralist" period, which was primarily concerned with (a) constructing tests which tested knowledge of discrete linguistic structures and rules, and (b) doing so with demonstrable statistical reliability and validity. More recently, however, there has been a reaction against this approach resulting in what Spolsky has termed the "integrative-sociolinguistic" approach to language testing which, while not discounting the importance of psychometric reliability and validity, puts a major emphasis on testing language as a functional, communicative tool as used in genuine communicative settings.

It is of particular interest to consider the dictation procedure as a language testing method from the perspective of these three different approaches to language testing. While the practice of having students

listen to and write down second language passages appears to have been fairly widely used as both a teaching and testing technique before the 1970s, it was generally later ignored by the true "psychometric-structuralists." Lado (1961, p. 34) describes dictation as a poor measure of language proficiency since both the words and their order is given by the examiner and, since the context of the passage may help the recognition of words which might not be recognized in isolation. Recently, however, the integrative-sociolinguistic approach to language testing has revived the use of dictation which is now seen by many as a convenient and valid language testing procedure which provides a useful measure of general language proficiency for those students who are familiar with the written form of the language.

Much of the impetus for the revival of dictation as a language testing procedure has come from the work of John Oller and his associates who have argued convincingly on both theoretical and empirical grounds for the convenience and validity of the dictation test (Oller, 1972, 1979; Oller & Streiff, 1975). Oller (1979, pp. 16-33) conceptualizes language proficiency as a "pragmatic expectancy grammer", i.e., a system of knowledge and rules which allow one to predict the form of language as it is being heard or read which permits comprehension as a constructive (or active) cognitive process (see Neisser, 1967, Clark & Clark, 1977, and van Dijk & Kintsch, 1983, for detailed theoretical considerations of language comprehension as a constructive, predictive process). This view of language proficiency is supported by a number of empirical studies (see Clark & Clark, 1977, pp. 210-215) which have demonstrated that (a) speech perception is an active process which requires the knowledge and use of top-down contextual constraints, and (b) the accuracy of recall of audi-

torily presented sentences similarly depends on knowledge of the lexical, syntactic, and semantic systems of the language. Thus, one of the very reasons for which Lado criticized dictation (i.e., it provides context which makes it easier to identify individual words) may be considered now to be dictation's most important characteristic as a language testing procedure since it is sensitive to one's integrative knowledge of the phonological, syntactic, and semantic systems of the language which permits its anticipation enabling both comprehension and production.

There also appear to be a number of practical reasons for the renewed popularity of the dictation procedure as a measure of second language proficiency. A dictation test is relatively easy to construct, requiring only the location of a passage of appropriate difficulty and style for the students to be tested and its division into segments (of usually 7 to 12 words) for presentation. It is therefore considerably easier to construct than multiple-choice tests (see Oller, 1979, Chap. 9) and very adaptable to the needs of individual classes. Thus, it is possible to create a dictation test with relative ease using an expository or narrative text or dialogue in either a formal or informal speech register at an appropriate level of difficulty (in terms of syntactic structure and vocabulary) including appropriate content. This adaptability of the dictation procedure gives it a number of advantages over available standardized language tests, particularly where formative evaluations of students' progress are desired and where specialized language skills are emphasized (e.g., the ability to read and write scientific articles in a specific technical field).

Nevertheless, there are a number of factors which limit the usefulness of the dictation procedure. Among these are:

1. In choosing a text for a dictation passage, there is no simple formula for deciding how difficult the text should be. This is of particular concern when a group of students representing a wide range of second language proficiency is to be tested.

2. While the usual procedure for scoring dictation involves subtracting one point for each insertion, deletion, permutation, and substitution at the word level, there is no clear theoretical or empirical basis for this particular weighting of all types of errors. Also, since total dictation scores which are equal may represent quite different patterns of responses, total test scores may not be easily comparable among examinees. For example, a score of 70 on a dictation test of 100 words may indicate quite different levels of language proficiency depending on whether missed points are primarily due to (a) omitted or inserted content words (e.g., nouns, verbs) which seriously affect the comprehensibility of the passage (and therefore would seem to indicate poor comprehension of the passage by the examinee) or (b) omitted or inserted functors (e.g., articles, conjunctions, prepositions) which are less important to the meaning of the text.

3. Although a dictation test is relatively easy to construct and administer, it requires considerably more time and care to score than most other tests requiring written responses (e.g., multiple-choice or cloze tests) if each individual word is to be scored.

4. The dictation procedure is limited to measuring listening comprehension and therefore cannot be used to assess language proficiency via the modality of reading.

Cziko (1982) felt that many of these shortcomings of the dictation procedure for measuring second language proficiency could be eliminated by making some basic changes to the way in which dictation is normally

administered and scored and by developing an analogous testing procedure which involved reading instead of listening as used in the dictation procedure. The principal changes to the dictation procedure involved presenting segments of the test text at widely varying lengths, from 2 to 21 words, and scoring each segment as a single item (right or wrong) instead of scoring each individual word. Cziko's major findings (as they relate to the four limitations of the traditional dictation procedure described above) were:

1. Varying the length of segments was effective in manipulating their difficulty resulting in a dictation test with a wide range of item difficulties appropriate for testing students possessing a wide range of language proficiency.

2. Awarding one point for each correct segment resulting in scores based on relatively fair items with surprisingly high reliability and validity. In addition, the procedure resulting in a Guttman scale of high reproducibility and scalability so that any given total score presented with few exceptions the same pattern of responses to each individual item (segment).

3. Scoring by segment was found to be three to four times faster than the conventional word by word scoring procedure.

4. The analogous test involving reading and writing (called a "copytest") administered to a smaller group of students did not have comparably high reproducibility or scalability. No analysis of its reliability or validity was undertaken.

Since one of the primary purposes of Cziko's (1981) study was to investigate whether this modification of the dictation procedure could result in a unidimensional, cumulative scale of language proficiency using Guttman

scalogram analysis, it should be mentioned here that Mokken (1971) has noted a number of problems associated with the use of the indices of scalability most often used to evaluate Guttman scales and has demonstrated that the index of test homogeneity ($\underline{H}$) proposed by Loevinger (1947, 1948) serves as a clearly better criterion of scalability. Also, Mokken (1971) and Mokken and Lewis (1982) have described a new index ($\underline{H}_i$) which is useful in evaluating the homogeneity and scalability of individual items within a given scale of items.

The purpose of the present study was to replicate the findings of Cziko (1981) that the modifications to the standard dictation test described above provide a practical and convenient procedure for obtaining reliable and valid short scales of language proficiency involving listening and reading. Unlike the previous study, however, three different approaches were used to analyze the resulting scales of language proficiency. These three approaches included (a) classical psychometric procedures for item analysis and reliability estimation, (b) a one-parameter latent trait (Rasch) model, and (c) a nonparametric scaling approach similar to Guttman's (1947, 1950) and Loevinger's (1947, 1948) concepts of a unidimensional, cumulative, and homogeneous scale which has been further refined by Mokken (1971) and Mokken and Lewis (1982).

Method

Subjects

A total of 67 students representing four levels of proficiency in English took part in this study. The beginning group (Group BEG, 13 students) and the intermediate group (Group INT, 12 students) were foreign adults and young adults studying at the Intensive English Institute (IEI) of the University of Illinois at Urbana-Champaign. Group BEG had

scored at the lowest level on the Illinois English Placement Test (IEPT) of all IEI students while Group INT had scored at a higher, intermediate level. Neither Group BEG nor Group INT was enrolled in regular university courses. The advanced group (Group ADV) were 25 foreign students enrolled in the University of Illinois who had scored high enough on the Test of English as a Foreign Language (TOEFL) to be admitted to their desired program of study but not high enough to be exempt from courses in English as a second language (ESL). A group of native English speakers (Group NS, 17 subjects) was selected from American undergraduates enrolled in an English rhetoric course. Thus, these four groups of subjects represented an extremely broad range of English proficiency, varying from extremely limited (Group BEG) to educated native-speaker competence (Group NS).

Materials

The text used for both the dictation and copytest was an adapted version of the introductory paragraph to the article entitled "Our Disappearing Wildlife" taken from a reader for intermediate and advanced ESL students (Lugton, 1978, p. 221). After pilot testing the passage as a dictation test with a small number of students comparable to students of Group INT, it was revised so that words which appeared to be too difficult were either omitted or changed to more common English synonyms.

Since we wanted to create a set of items representing a wide range of difficulty, a technique was sought to manipulate the difficulty of the segments of the passage which were to be used as test items. While it was recognized that a large number of factors including segment length, vocabulary difficulty, syntactic complexity, and speed of presentation would all likely influence the ease with which a prose segment could be compre-

hended and recalled, it was apparent that manipulating the length of the segments was by far the most convenient way of providing a set of items with widely ranging difficulty levels. Thus, the test version of the passage consisted of 105 words divided into 13 segments of generally increasing length ranging from 2 words (first segment) to 19 words (last segment; see Appendix). These 13 segments were formed by dividing the text at the natural division points provided by phrase, clause, or sentence boundaries.

Procedure

The administration of each test involved three complete presentations of the test passage, either auditorily via an audiotape recording (for the dictation) or visually via typed transparencies on an overhead projector (for the copytest). The entire testing session lasted approximately 15 minutes for each test and included: (a) test instructions, (b) the first presentation of the test passage during which the entire passage was presented to the subjects without interruption, (c) the second presentation of the test passage divided into segments which included pauses at the end of each of the 13 segments to allow the students time to write what they had heard or read, (d) the third presentation of the test passage with pauses at the end of each of the seven sentences of the passage to allow the students to check and correct what they had written, and (e) a 60-second pause after the third present.tion for final corrections.

For all dictation presentations of the test passage, the passage was read at a speed considered normal for a careful oral reading of a written text. For all copytest presentations, the time taken to read the text for the dictation test (or portions of the text for the second and third presentations) was used as the visual presentation time for the text and

portions of the text. The length of pauses used for the second and third presentations were determined by estimating the time required by the students to write and correct their work. For the second presentation, the length of each pause in seconds after each segment was determined by dividing the number of letters in the segment being written by 2. For the third presentation, the length of each pause in seconds was determined by dividing the number of letters in the sentence being written by 5.

Since the same passage was used for both the dictation and copytest and since the same students took both tests, the order of administration of the tests was counterbalanced with approximately half of the students (35) taking the dictation test first and the remaining half (32) taking the copy-test first. For all students, there was an interval of three days between the administration of the two tests. Both tests were administered during a regular ESL class period to each of the four groups separately.

Scoring

For both the dictation and copytest, each of the 13 segments was considered one item. Students were given one point for each segment written without error (including spelling errors) with 13 as the maximum score for each test. While a number of researchers have recommended a less strict scoring procedure which allows for some spelling errors and gives at least partial credit to responses which retain the basic meaning of the test passage (see Oller, 1979; Savignon, 1982), a less strict scoring procedure is relatively time consuming and likely to be less reliable since it requires a subjective judgment on the part of the scorer as to what errors should be given credit. Also, Cziko (1982) found that the exact-segment scoring criterion used for a dictation test did result in test scores with high reliability and validity and that the dictation test segments

scored in this way formed a Guttman scale of high reproducibility and scalability.

Results

Order Effects

There was a small but quite consistent order effect for the two tests. This was indicated by the finding that regardless of group, those students who took the dictation after having taken the copytest did better on the dictation than those students who took the dictation first. These differences in group means (out of a maximum possible score of 13) were .98, 1.17, 1.35, and .12 for Groups BEG, INT, ADV, and NS, respectively. The same was also generally true of the copytest with the exception of Group ADV for whom the order of administration had virtually no effect. Order differences in group means for the copytest were .42, .67, -.04, and .70 for Groups BEG, INT, ADV, and NS, respectively. However, none of the above differences was statistically significant ($p > .05$) when tested using the t statistic and the directional alternative hypothesis that group means would be higher for those students who took a given test second. Therefore, all further analysis were done without regard to order of administration.

Item and Scale Analyses

Three different approaches were used to analyze the item and scale characteristics of the dictation and copytest (see Table 1). Since Rasch analysis requires the exclusion of students receiving zero or perfect scores on a test, these extreme subjects were excluded from the analyses of all three approaches so that all results would be based on the same students for each of the two tests. Thus, data from 47 and 56 students were

included in the following analyses of the dictation and copytest, respectively.

First, standard psychometric indices were computed for each item using the reliability procedure of SPSS (Hull & Nie, 1981). The indices included item easiness ($\underline{p}$, i.e., the proportion of students passing each item), the corrected item-total point-biserial correlation coefficient ($\underline{r}_{pb}$), and the value of Cronbach's α for the entire scale deleting a given item. These analyses indicated that both tests included items of widely ranging easiness ($.09 \leq \underline{p} \leq .83$ for the dictation, $.11 \leq \underline{p} \leq .98$ for the copytest) with the first three items of each test having noticeably lower values of $\underline{r}_{pb}$ than the remaining items of each test. It was also found that while the values of "α if deleted" were highest for these first three items of each test, variation in these values across items was very small.

Insert Table 1 about here

In the second approach to analyzing these scales, the Rasch model (a one-parameter logistic latent trait model) was used to fit the data generated by each of the two tests using Wright and Mead's (1977) BICAL computer program. A one-parameter item response model was considered appropriate since guessing was not a factor influencing performance on the tests (all responses were supplied by the students, not selected), the number of subjects was relatively small, and previous research with similar scales (Cziko, 1981) revealed that they resembled Guttman scales with each item having similar high discriminatory power. Dividing the examinees into two groups (24 low scorers and 23 high for the dictation; 30 low and 26 high for the copytest), total fit and discrimination indices were computed

for each item (see Table 1). Except for the third copytest item, all items provided quite acceptable total fit indices which were well within three standard error units of the expected total fit values of unity (standard errors of expected total fit were .21 and .19 for the dictation and copytest, respectively). The Rasch analyses also indicated that with five exceptions (out of a total of 26 items), the discriminating power of items in both the dictation and copytest were comparable. With a value of unity indicating that an item's observed characteristic curve is equal in steepness to the best fitting logistic curve for all items, the first and third items of the dictation as well as the third copytest item were found to have relatively flat curves while the fifth item of the dictation and the sixth and eighth items of the copytest were found to have relatively steeper curves and consequently higher discriminating power than the other items of their respective scales.

Finally, the $\underline{H}_i$ statistic formulated by Mokken (1971) was calculated for each item using Cziko's (1984) computer program. This statistic is similar to Loevinger's $\underline{H}$ (Loevinger, 1947, 1948) in that it provides an indication of scale homogeneity and scalability. However, whereas Loevinger's $\underline{H}$ can only be used to evaluate the homogeneity or scalability of a complete set of items, Mokken's $\underline{H}_i$ provides a way of evaluating each item's contribution to the homogeneity or scalability of the scale of which it is a part. Using the criteria proposed by Mokken (p. 185) of considering values of .5 or above as evidence of strong scalability, .4 to .5 as evidence of a medium scalability, and .3 to .4 indicating weak scalability, we notice 19 "strong" items, 1 "medium" item, 3 "weak" items and 2 nonscale items with $\underline{H}_i$ of less than .3. Again, all weak or nonscale items were found among the first three items of each test.

In comparing the above three approaches to scale and item analysis, all three showed a high degree of convergence in signalling items 1, 2, and 3 of the dictation and items 1 and 3 of the copytest as items with a relatively poor fit to the scale defined by the other items. However, while the item-total correlation and scalability for item 2 of the dictation were relatively low (r_{pb} = .38, H_i = .34), this item nevertheless had close to expected fit and discrimination indices according to the Rasch analysis. Also, while item 5 of the dictation and item 8 of the copytest showed a much steeper discrimination curve than other items in their respective scales, all other indices of fit for these two items appeared quite acceptable.

Indices of the reliability and homogeneity of the dictation and copytest are given in Table 2. In spite of the fact that each test consisted of only 13 items and that students with extreme scores were excluded from these analyses, all estimates of psychometric reliablity were in the range of .82 to .90. In additon, the dictation and copytest were found to have H values of .50 and .58, respectively, indicating that they comprised what could be considered strong homogeneous scales (Mokken & Lewis, 1982, p. 422).

Insert Table 2 about here

Validity

Two principal techniques were employed to assess the construct validity of the two language proficiency measures. These included (a) comparing the mean dictation and copytest scores of the four groups of students, and (b) examining the correlations of the dictation and copytest scores with other tests of English reading and listening comprehension.

A summary of the performance of the four groups on the dictation and copytest is given in Table 3. For both measures, the relative magnitudes of all group means were as predicted with Group BEG scoring lowest, followed in order of increasing mean scores by Groups INT, ADV, and NS. Differences in means between adjacent groups were shown to be quite large when divided by the pooled standard deviation of test scores for all four groups. The resulting effect size (ES) was well above unity for each comparison with the largest values obtained when comparing Groups ADV and NS. Confidence intervals of the difference between adjacent group means (C = .95) ranged from a lower limit of .77 (Groups INT and ADV on the copytest) to 8.01 (Groups ADV and NS on the dictation). These analyses provide evidence of the validity of the dictation and copytest in that the ordering of the group means was consistent with the ordering that a valid test of English proficiency would be expected to produce and differences between adjacent group means were large and statistically significant. In addition, all but one student of Group NS scored 10 or above on each test whereas the majority of students in Groups BEG and INT scored 5 or below on each test.

Insert Table 3 about here

Pearson product-moment intercorrelations were computed among the dictation and copytest total scores, the log ability scores of students with nonextreme dictation and copytest scores, the subparts and total of the IEPT (dictation, structure, and cloze tests), and the subparts and total of the TOEFL (listening comprehension, structure, and reading comprehension tests). The upper triangle of Table 4 gives correlation coefficients using

all available data. Since no students from Groups ADV or NS had recently taken the IEPT or TOEFL, all correlations involving these tests were based on relatively small numbers of students (18 to 25). Also, since the correlation coefficients in the upper triangle of Table 4 are based on different numbers and subgroups of students, correlation coefficients based on data from the same set of 18 students from Groups BEG and INT who took all tests listed in the table were also computed and are presented in the lower triangle.

Among these correlations, of particular interest is that both the dictation raw scores and dictation log ability scores had moderately high correlations with the TOEFL listening comprehension test (.84 and .82 on the upper and .74 and .82 on the lower triangle, respectively). Similarly, the copytest raw scores and the copytest log ability scores correlated quite well with the TOEFL reading comprehension test (.68 and .67 on the upper and .65 and .70 on the lower triangle, respectively). Also, while the dictation and copytest used completely different methods for presenting the test text (auditory vs. visual), correlations between both raw and log ablity scores were quite high when based on all available data (.89 and .79 for raw scores and log ability scores, respectively). Finally, on the lower triangle (where the intercorrelations can be more meaningfully compared since they are based on the same group of students), the log ability scores of the dictation and copytest had, with only one exception out of 20 comparisons, uniformly higher correlations with the remaining language tests than did the simple total (raw) scores of the dictation and copytest. While these differences were not great (ranging from .03 to .08), they do suggest that the non-linear transformations of total dictation and copytest scores provided by the Rasch analysis were better predictors of perfor-

mance on other measures of langauge proficiency than were the simple total scores.

At this point it seems appropriate to address two concerns arising from the nature of these two novel testing procedures. Since Table 1 shows a clear relationship between item length and difficulty for the two scales, it may seem that longer items were in general more difficult simply because they presented more opportunities to err than the shorter items. Also, since the items gradually increased in length throughout each test culminating in a segment of 19 words, these tests may in some respects appear more like tests of short-term memory than of language proficiency. While both of these concerns have some validity, it is nevertheless the case that the longer items did very well in discriminating Group NS from all of the ESL students. For example, on the dictation test all 17 Group NS students passed the last item while only one ESL student from Group ADV did so while on the copytest 15 out of 17 Group NS students passed the last item while only 10 out of 50 ESL students did so (8 from Group ADV, 2 from Group INT, and 0 from Group BEG). Therefore, unless there is some reason to believe that native English-speaking American students have uniformly better short-term memories than foreign students, it appears more reasonable to conclude that it is the different levels of language knowledge represented in the sample that is responsible for the variation in test scores (see Table 3). It is this knowledge which is necessary for the comprehension and "chunking" of the words in each item which permits their retention in short-term memory (see Miller, 1956). Also, while it cannot be denied that longer items present more opportunities for error, this is also likely the case for most mental tests where more difficult items (e.g., reading test items requiring the integration of

many pieces of information as well as inferencing skills; mathematics problems requiring many computational steps) generally present many more opportunities for error than easier items. Thus it could be argued (as the preceding validity analyses suggest) that the longer items are more difficult for the "right" reason in that they require exactly the kind of mental processing which is made possible only by knowledge of the English language and which is in fact required in all forms of comprehending and producing language.

In summary, the group mean differences and intercorrelations reported in this section suggest that the dictation and copytest are valid measures of language proficiency and that the Rasch log ability transformations of the total dictation and copytest scores have slightly higher validity as measures of language proficiency than do the raw total scores of these two tests.

Re-analysis of Cziko (1982) Data

Since the results reported above are based on relatively small numbers of students and since the dictation and copytest used the same text segmented in identical ways, the dictation and copytest data reported by Cziko (1982) were re-analyzed in an attempt to replicate these findings using the same methods of scale analysis. These data were collected using a dictation and copytest based on a different text than the one used above consisting of 14 items ranging in length from 2 to 21 words. A total of 102 students were administered the dictation and a smaller group of 34 students took the copytest. As above, these students represented beginning to native-speaker proficency in English. Excluding all students with either perfect or zero test scores from the analyses left 87 and 33 students for the dictation and copytest, respectively.

Insert Tables 5 and 6 about here

The results of the analyses of these dictation and copytest items and scales are presented in Tables 5 and 6. As can be seen on Table 5, all 14 items of the dictation and all but the second item of the copytest appeared to have acceptable indices of Rasch fit and discrimination as well as acceptable scalability as indicated by $\underline{H}_i$ (the standard errors of total expected Rasch fit were .21 and .19 for the dictation and copytest, respectively). While the corrected item-total point-biserial correlations (r_{pb}) were lower for extreme items, the other indices indicate that except for the second copytest item, these extreme items were nonetheless homogeneous within their respective scales. As shown on Table 6, both tests had estimates of psychometric reliability ranging from .85 to .93 with high $\underline{H}$ values of .76 and .61 for the dictation and copytest, respectively, indicating that both tests formed strong cumulative, homogeneous scales. With respect to the construct validity of the dictation test, Cziko (1981) reported that group mean differences and correlations with other measures of language proficiency (ranging from .75 to .86) supported its validity as measure of language proficiency.

Discussion

The results of this research have provided evidence that relatively short scales can be constructed to provide useful measures of language proficiency. It appears that such scales can be easily constructed by manipulating the length of segments of coherent text, presenting the segments either auditorily (as for the dictation) or visually (as for the

copytest), and requiring the examinees to write down what they recall after the presentation of each segment.

The use of homogeneous, cumulative scales to measure language proficiency has a number of both theoretical and practical advantages over most other language testing techniques. First, the homogeneity and cumulativeness of a set of such items can be considered evidence of its unidimensionality, a quality which is important for all measures of ability and yet has been found to be quite difficult to reliably assess using even sophisticated factor analytic and latent trait procedures (see Hambleton, 1983).

Second, since the items of an ability test can only be cumulative if the scale includes items from all along the difficulty continuum from very easy to very difficult, such a test can be used for students representing a very broad range of language proficiency, ranging from very poor to native-speaker proficiency. This cumulativeness of the items also assures that an individual's total score is a good predictor of responses to each of the individual items. This makes test scores more directly comparable and meaningful since two individuals obtaining same total test score on a cumulative scale will have a similar pattern of responses to individual test items. Cumulativeness also makes it possible to examine the response patterns of individuals for evidence of inattentiveness to the test or cheating. Such behavior would be indicated by a response pattern characterized by the failing of easy items and the passing of more difficult items (see Harnisch, 1983, for a detailed discussion of unusual item response patterns, how they can be quantified, and their implications for testing and instruction).

Also, these scales are amenable to Rasch analysis since guessing is not a factor and since with few exceptions items were found to have consistently high discriminatory power. The log ability scores provided by Rasch analysis correlated in general more highly with other measures of language proficiency than did the raw dictation and copytest scores. Since the estimation of only one parameter for each item is less demanding in terms of the number of examinees required than two- and three-parameter item response models, the Rasch method as employed here can be used in settings where relatively small numbers of students would make two- and three-parameter approaches inappropriate.

While the three approaches used to analyze the characteristics of these scales tended to converge in signalling the same items as suspect, the corrected point-biserial item-total correlations were influenced by the centrality of test items (always giving lower coefficients to very easy or very hard items) while the Rasch and Mokken indicators were not so influenced. Since point-biserial correlations are so influenced by item difficulty, the Rasch and/or Mokken indices as used here appear more appropriate for analyzing items included in a scale of items with widely ranging difficulties.

Among the practical advantages of the language testing procedures investigated in this research are the ease and speed with which these tests can be scored in comparison to traditional scoring methods which require attention to each individual word of the test passage. This segment scoring procedure is significantly faster than the usual dictation scoring procedure which treats each word of the test passage as separate item. This feature, along with the high reliability and validity of the procedures studied in this research, resulted in a decision to replace the traditionally

designed dictation test of the IEPT with the dictation test used in Cziko's (1981) research.

Finally, since these language testing measures are based on coherent text, they are relatively easy to construct. They also allow the flexibility of using either written texts or dialogues, depending on the type of language one wishes to test. While such text reconstruction tasks have in the past employed primarily written expository and narrative texts, there is no apparent reason why texts based on naturally occurring oral language could not be used as well. Thus, it appears possible to use such techniques to test proficiency in a wide variety of styles, dialects, and registers of the target language.

Even though both of the language testing techniques investigated in this research involved writing as the response, there is no reason why oral production could not be used as the response mode should it be desirable not to involve writing. While oral language production in response to auditorily presented language has been used in tests of elicited imitation (see, e.g., Swain, Dumas, & Naiman, 1974), the authors are not familiar with tests of second leanguage proficiency that have used oral reading tasks. Research is needed to determine whether such language testing procedures, modified according to the techniques used in this study, would have the same desirable characteristics as the dictation and copytest procedures investigated here. If this is the case, we would then have four powerful and practical language testing procedures for measuring language proficiency which involve either the auditory or visual presentation of language and either writing or oral production as response modes.

The present research has demonstrated that relatively short, cumulative scales can provide reliable and valid measures of language proficiency. It is hoped that this finding will encourage measurement specialists to investigate ways in which such scales can be used to measure other cognitive variables and to no longer consider these measurement techniques appropriate only for the measurement of affective variables in the way that Guttman scales have been primarily used.

References

Clark, H. H., & Clark, E. V. (1977). Psychology and language. New York: Harcourt Brace Jovanovich.

Cziko, G. A. (1981). Psychometric and edumetric approaches to language testing: Implications and applications Applied Linguistics, 2, 28-44.

Cziko, G. A. (1982). Improving the psychometric, criterion-referenced and practical qualities of integrative language tests. TESOL Quarterly, 16 (3), 367-379

Cziko, G. A. (1984). An improvement over Guttman scalogram analysis: A computer program for evaluating cumulative, nonparametric scales of dichotomous items. Educational and Psychological Measurement, 44, 157-161.

Guttman, L. (1947). The Cornell technique for scale and intensity analysis Educational and Psychological Measurement, 7, 247-280.

Guttman, L. (1950). The basis for scalogram analysis. In S. Stouffer, L. Guttman, E. Suchman, P. Lazarsfeld, & J. Clausen (Eds.), Measurement and Prediction. Princeton, N. J.: Princeton University Press.

Hambleton, R. K. (1983, April). Assessing dimensionality of a set of test items. Paper presented at the meeting of the American Educational Research Association, Montreal.

Harnisch, D. L. (1983). Item response patterns: Applications for educational practice. Journal of Educational Measurement, 20, 191-206.

Hull, C. H., & Nie, N. H. (1981). SPSS Update 7-9. New York: McGraw-Hill.

Lado, R. (1961). Language testing. New York: McGraw-Hill.

Loevinger, J. (1947). A systematic approach to the construction of and evaluation of tests of ability. Psychological Monographs, 61 (4).

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. Psychological Bulletin, 45, 507-530.

Lugton, R. (1978). American topics. Englewood Cliffs, N.J.: Prentice-Hall.

Miller, G. A. (1956). The magical number seven plus or minus one or two. Psychological Review, 63, 81-97.

Mokken, R. J. (1971). A theory and procedure of scale analysis. The Hague: Mouton.

Mokken, R. J., & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. Applied Psychological Measurement, 6, 417-430.

Neisser, U. (1967). Cognitive psychology. New York: Appleton-Century-Crofts.

Oller, J. W., Jr. (1972). Dictation as a test of ESL proficiency. In H. B. Allen & R. N. Campbell (Eds.), Teaching English as a second language. New York: McGraw-Hill.

Oller, J. W., Jr. (1979). Language tests at school. London: Longman.

Oller, J. W., Jr., & Streiff, V. (1975). Dictation: A test of grammar based expectancies. In R. L. Jones & B. Spolsky (Eds.), Testing language proficiency. Arlington, VA: Center for Applied Linguistics.

Savignon, S. J. (1982). Dictation as a measure of communicative competence in French as a second language. Language Learning, 32, 33-51.

Spolsky, B. (1978). Introduction: Linguists and language testers. In B. Spolsky (Ed.), Approaches to language testing. Arlington, VA: Center for Applied Linguistics.

Swain, M., Dumas, G., & Naiman, N. (1974). Alternatives to spontaneous speech: Elicited translation and imitation as indicators of second language competence. Working Papers in Bilingualism, 3, 76-90. (ERIC Document Reproduction Service No. ED 123 872)

Wright, B. D., & Mead, R. J. (1977). BICAL: Calibrating items and scales with the Rasch model (Research Memorandum No. 23). Chicago: University of Chicago, Department of Education.

Author Notes

Correspondence concerning this article should be sent to Gary A. Cziko, University of Illinois, Department of Educational Psychology, 1310 S. Sixth St., Champaign, IL 61820.

The current address of Nien-Hsuan Jennifer Lin is 132 Rancho Drive #275, San Jose, CA 95111.

Table 1

Characteristics[a] of Dictation and Copytest Items

Sequence	Length	p	r_{pb}	α	Difficulty	Fit	Discrimination	H_i
				Dictation (n=47)				
1	2	.83	.19	.83	-3.05	1.23	.70	.33
2	3	.45	.38	.82	-.55	1.21	.93	.34
3	3	.91	.13	.83	-3.99	1.35	.70	.31
4	3	.62	.49	.81	-1.58	.68	1.06	.53
5	4	.49	.60	.80	-.81	.59	1.48	.54
6	6	.38	.56	.80	-.13	.73	1.05	.50
7	5	.23	.50	.81	1.04	.86	1.16	.51
8	7	.43	.59	.80	-.41	.63	1.03	.51
9	10	.43	.54	.80	-.41	.81	1.03	.47
10	11	.09	.45	.81	3.00	.63	1.04	.64
11	15	.13	.59	.80	2.28	.29	1.06	.69
12	17	.15	.51	.81	1.99	.58	1.08	.58
13	19	.11	.48	.81	2.62	.55	1.05	.61
				Copytest (n=56)				
1	2	.98	.04	.85	-5.15	1.75	.99	.14
2	3	.91	.25	.85	-3.31	.88	1.04	.53
3	3	.95	-.11	.86	-3.96	53.16	.02	-.28
4	3	.63	.60	.83	-.64	.61	1.01	.62
5	4	.55	.58	.83	-.11	.70	1.03	.54
6	6	.54	.80	.81	.03	.31	1.42	.71
7	5	.52	.50	.84	.16	1.16	.84	.46
8	7	.52	.72	.82	.16	.48	1.43	.64
9	10	.41	.58	.83	1.00	1.05	.97	.56
10	11	.23	.60	.83	2.60	1.15	1.03	.69
11	15	.18	.62	.83	3.19	.24	1.13	.80
12	17	.11	.41	.84	4.11	.39	1.10	.68
13	19	.38	.49	.84	1.91	1.13	.98	.52

Note. Data from students obtaining zero or perfect scores were excluded from these analyes.

[a]Definitions of these item characteristics are: sequence = order of item in passage; length = number of words in item; p = proportion of students passing item; r_{pb} = corrected item-total point-biserial correlation; α = internal consistency of test with item deleted; fit = Rasch total mean-square fit; discrimination = Rasch item discrimination; H_i = Mokken index of item homogeneity.

Table 2

Reliability and Homogeneity of Dictation and Copytest Scales

Characteristic	Dictation	Copytest
Cronbach's α	.82	.84
Spearman-Brown split-half reliability	.86	.86
Guttman split-half reliability	.86	.86
Guttman largest λ	.86	.90
Loevinger's H	.50	.58

Note. The data of the same students included in Table 1 were included in these analyses.

Table 3

Dictation and Copytest Results

Group	n	M	SD	ES[a]	Estimate of difference between means (C = .95)[b] Lower limit	Upper limit
			Dictation			
BEG	13	1.00	1.53			
				1.31	1.21	3.63
INT	12	3.42	1.38			
				1.33	.86	4.06
ADV	25	5.88	2.54			
				3.63	5.41	8.01
NS	17	12.59	.87			
			Copytest			
BEG	13	2.54	1.66			
				1.22	.97	3.95
INT	12	5.00	1.95			
				1.23	.77	4.19
ADV	25	7.48	2.58			
				2.31	3.29	5.99
NS	17	12.12	1.11			

Note. Data from all students were included in these analyses.

[a]ES for adjacent group means was calculated by subtracting the mean of the less proficient group from the mean of the more proficient group and dividing this difference by the pooled standard deviation of test scores for all four groups.

[b]These estimated limits are for adjacent means shown on rows immediately above and below the row on which the limits are given.

Table 4

Intercorrelations Among Measures of Language Proficiency

		Listening tests				Reading tests							
	Test	1	2	3	4	5	6	7	8	9	10	11	12
1.	Dictation total	--	.99(47)	.49(25)	.84(23)	.89(67)	.83(56)	.77(25)	.71(25)	.61(23)	.56(23)	.77(25)	.74(23)
2.	Dictation log ability	.98	--	.27(18)	.82(18)	.81(47)	.79(46)	.62(18)	.77(18)	.46(18)	.59(18)	.68(18)	.68(18)
3.	IEPT dictation	.23	.27	--	.64(23)	.47(25)	.29(22)	.75(25)	.55(25)	.62(23)	.58(23)	.81(25)	.68(23)
4.	TOEFL listening comprehension	.74	.82	.49	--	.45(23)	.32(22)	.90(23)	.71(23)	.69(23)	.74(23)	.89(23)	.89(23)
5.	Copytest total	.30	.34	.40	.35	--	.99(56)	.56(25)	.59(25)	.35(23)	.68(23)	.61(25)	.56(23)
6.	Copytest log ability	.33	.37	.38	.39	.99	--	.43(22)	.57(22)	.37(22)	.67(22)	.50(22)	.52(22)
7.	IEPT structure	.55	.62	.65	.84	.52	.57	--	.73(25)	.84(23)	.80(23)	.97(25)	.94(23)
8.	IEPT cloze	.74	.77	.34	.63	.55	.63	.68	--	.65(23)	.78(23)	.85(25)	.80(23)
9.	TOEFL structure	.40	.46	.58	.62	.42	.47	.81	.66	--	.73(23)	.83(23)	.89(23)
10.	TOEFL reading comprehension	.52	.59	.50	.74	.65	.70	.87	.74	.80	--	.84(23)	.93(23)
11.	IEPT total	.63	.68	.71	.81	.58	.64	.96	.82	.83	.87	--	.94(23)
12	TOEFL total	.61	.68	.58	.85	.55	.60	.93	.76	.89	.96	.93	--

Note. Each coefficient above the main diagonal includes all students with non-missing data for the two tests (the number of students for each coefficient is given in parentheses). Each coefficient below the diagonal includes the same 18 students from Group BEG and INT for whom test scores on all tests were availble. All correlation coefficients greater than .39 were significantly greater than zero ($p < .05$).

Table 5

Re-Analysis of Characteristics of Dictation and Copytest Items from Data Collected by Cziko (1982)

Sequence	Length	p	r_{pb}	α	Difficulty	Fit	Discrimination	H_i
				Dictation (n=87)				
1	2	.94	.18	.91	-7.69	.49	1.13	.68
2	4	.80	.35	.91	-5.70	.61	1.14	.78
3	4	.68	.51	.90	-4.25	.34	1.14	.89
4	6	.39	.68	.89	-.99	.88	1.03	.79
5	5	.33	.75	.89	-.34	.42	1.03	.81
6	8	.25	.52	.90	.68	1.36	.95	.53
7	8	.25	.78	.89	.68	.32	1.03	.78
8	7	.22	.77	.89	1.19	.36	1.03	.77
9	10	.21	.76	.89	1.37	.38	1.03	.76
10	10	.20	.81	.89	1.56	.22	1.03	.81
11	13	.18	.72	.89	1.75	.33	1.03	.74
12	14	.11	.60	.90	3.02	.31	1.04	.76
13	18	.10	.62	.90	3.25	.17	1.04	.83
14	21	.02	.27	.90	5.47	.19	1.02	.76
				Copytest (n=33)				
1	2	.94	.43	.84	-4.62	.13	1.08	.84
2	4	.85	.14	.86	-2.89	3.73	-.14	.19
3	4	.88	.62	.83	-3.38	.12	1.08	.87
4	6	.73	.59	.83	-1.48	.74	1.00	.62
5	5	.76	.52	.84	-1.77	.99	.93	.57
6	8	.39	.57	.83	1.03	.64	.87	.61
7	8	.61	.70	.82	-.49	.44	1.33	.69
8	7	.67	.68	.83	-.96	.44	1.30	.68
9	10	.52	.72	.82	.17	.41	1.45	.71
10	10	.18	.40	.84	2.76	.65	1.09	.59
11	13	.42	.62	.83	.81	.56	1.20	.64
12	14	.27	.29	.85	1.94	1.56	.91	.37
13	18	.06	.33	.85	4.44	.16	1.04	.74
14	21	.06	.17	.85	4.44	1.72	1.04	.39

Note. Data from students obtaining zero or perfect scores were excluded from these analyses. See note of Table 1 for definitions of item characteristics.

Table 6

Re-Analysis of Reliability and Homogeneity of Dictation and Copytest Scales from Data Collected by Cziko (1982)

Characteristic	Dictation	Copytest
Cronbach's α	.90	.85
Spearman-Brown split-half reliability	.92	.89
Guttman split-half reliability	.92	.88
Guttman's largest λ	.93	.91
Loevinger's H	.76	.61

Note. The data of the same students included in Table 3 were included in these analyses.

Appendix A

Test Passage

Wild animals / used to wander / over our country / in uncounted numbers. / Today these animal populations / have decreased to a great extent. / Some animals have disappeared altogether, / destroyed by the advance of human civilization. / The same story can be told in the African continent, / once covered with big game such as elephant, buffalo, and antelope. / In Central and South America, where animals were once thought safe, they are now threatened. / In the last three centuries, over two hundred species of mammals, birds, and reptiles have become extinct. / Our wild animals are being swept from the land, the birds from the air, the fish from the sea.

Note. The boundaries of the 13 segments (items) of the test passage are indicated by ~~vertical~~ slanted lines.